# An Explainable AI model in the assessment of Multiple Sclerosis using clinical data and Brain MRI lesion texture features*

A. Nicolaou, *Member, IEEE*, M. Pantzaris, C. P. Loizou, *Senior Member, IEEE,*
A. Kakas, C. S. Pattichis, *Fellow, IEEE*

***Abstract*— Magnetic resonance imaging (MRI) is an essential visualizing tool in the diagnosis and monitoring of Multiple Sclerosis (MS) disease. However, the neurological examinations and the MRI assessments are insufficient to provide personalized treatment to the patients due to the complexity of the disease. This study implemented an explainable artificial intelligence (AI) model with embedded rules to assess MS disease evolution. Clinical data were used including demographic and neurological measurements. Texture features were extracted from manually delineated and normalized brain MRI lesions. Statistical analysis was employed to select the statistically significant texture features and clinical data. Different models using machine learning algorithms were implemented to differentiate the subjects diagnosed with relapsing-remitting MS (RRMS) from the subjects with progressive MS (PMS). Argumentation-based reasoning was performed by modifying the rules extracted from models with the best evaluation results. The findings indicated that the proposed explainable AI model can predict the clinical conditions of MS disease with high accuracy and provide transparent and understandable explanations with high fidelity. Future work will include further clinical data such as medications and investigate the correlation of the texture features and clinical data with the neurological impairment. The proposed model should also be evaluated on more MS subjects.**

***Clinical Relevance*— This method can assist clinical experts by providing explainable and interpretable diagnosis in the assessment of MS disease.**

## I. Introduction

Clinically, Multiple Sclerosis (MS) is marked by a wide range of symptoms stemming from localized inflammation in the central nervous system, which can manifest at different periods. These symptoms usually endure for several days to weeks, but occasionally persist for months, followed by complete or partial recovery. These periods are known as relapses [1]. On brain imaging, MS is detectable by the white matter (WM) lesions which can be visualized using magnetic resonance imaging (MRI) [2]. The relapses and the lesions can be used to estimate the disease activity [1], [2].

During the initial stages of the disease, relapses are the main clinical manifestation in most patients, a condition referred to as relapsing-remitting MS (RRMS). When the disease progresses, secondary-progressive MS (SPMS) may develop, characterized by independent progression of the disease apart from relapses, leading to a gradual decline in neurological function and increasing disability over time. In instances where the disease progresses without early relapses, it is termed primary-progressive MS (PPMS) [1].

Explainable artificial intelligence (AI) methods light up the black box nature of machine learning (ML) [3]-[5] and deep learning (DL) models [6], providing transparent and understandable explanations to assist the experts in MS diagnosis. More specifically, the following models were used: Shapley additive explanations (SHAP) [3]-[5] and local interpretable model-agnostic explanations (LIME) [4] were used in ML models, and layer-wise relevance propagation (LRP) [6], was used in DL models.

The objective of this study was to implement an explainable AI model with embedded rules to assess MS disease focused on clinical data and brain MRI lesion texture features.

## II. Methodology

The proposed methodology consisted of five main processing steps which are analyzed below.

### A. MRI acquisition and clinical data

A dataset of 87 MS subjects (34 males, and 53 females) was examined at different time points. MRI images of 66 RRMS and 21 progressive MS (PMS) including both SPMS and PPMS patients were obtained using five different MRI scanners and different sequences ($T_{1w}$, $T_{2w}$, and FLAIR). Clinical data were also investigated including demographic, and neurological measurements, such as functional system (FS) scores ranging from 0 (normal) to 6 (loss) [7], timed 25-foot walk (T25-FW), and 9-hole peg test (9-HPT) measures. More information can be found in Table I.

TABLE I. CLINICAL DATA SETS

| **Demographic** |
|---|
| Age, gender, duration of the disease (in years) |
| **Neurological measurements** |
| Functional systems (FS): Brainstem (Extraocular muscles, Nystagmus, Trigeminal, Facial, Hearing Loss, Dysarthria, Dysphagia, Slow Tongue RAM, Other Bulbar Signs), Visual, Pyramidal, Cerebellar, Sensory, Bowel and Bladder, Cerebral |
| Timed 25-foot walk (T25-FW): Trial 1, Trial 2 |
| 9-hole peg test (9-HPT): Dominant hand (Trial 1, Trial 2), Non-dominant hand (Trial 1, Trial 2) |

*Research partly supported by the University of Cyprus scholarship program awarded to the first author.

A. Nicolaou, A. Kakas and C. S. Pattichis are with the Department of Computer Science, University of Cyprus, Nicosia, Cyprus (e-mail: {nicolaou.andria, antonis, pattichi}@ucy.ac.cy).

M. Pantzaris is with the Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus (e-mail: pantzari@cing.ac.cy).

C. P. Loizou is with the Department of Electrical Engineering, Computer Engineering and Informatics, Cyprus University of Technology, Limassol, Cyprus (email: christos.loizou@cut.ac.cy).

### B. Preprocessing

The expert neurologist (co-author, M. Pantzaris) manually delineated the brain MS lesions in a blinded manner where the subject, the time of the exam, and the clinical findings could not be identified.

### C. Feature extraction and selection

Texture features were extracted from all the segmented MS lesions and were estimated by averaging the corresponding values for all lesions of each patient. The following selected features were extracted [8]: (1) first-order statistics (FOS): mean, variance, median, mode, skewness, kurtosis, energy, entropy; (2) spatial grey level dependence matrix (SGLDM): angular second moment (ASM), contrast, correlation, variance - sum of squares (SOSV), homogeneity - inverse difference moment (IDM), sum average, sum variance, sum entropy, entropy, difference variance, difference entropy; (3) neighborhood grey tone difference matrix (NGTDM): coarseness, contrast, busyness, complexity, strength.

A min-max scaler was used to normalize the features between the values 0.0 and 1.0, where a fixed number of 3 bins that has the same number of observations to each bin (quantile strategy) was defined. The bins were encoded using the ordinal method, where 0 refers to 'Low', 1 refers to 'Medium' and 2 refers to 'High'.

### D. Classification analysis

Classification modeling was developed to differentiate the subjects diagnosed with RRMS ($G_1$) from those with PMS ($G_2$) based on clinical data and brain MRI lesion texture features. The classification models were implemented in Python using the scikit-learn [9] and xgboost [10] libraries. Different classifiers were used, such as decision tree (DT), random forest (RF), gradient boosting (GB), extreme gradient boosting (XGB), k-nearest neighbors (kNN), gaussian naïve Bayes (NB), and support vector machine (SVM). As shown in Table II, data were split into a training group and an evaluation group, using 80% for the training and 20% for the evaluation set, and were re-arranged to have an equal size of the two classes on the evaluation set. The synthetic minority over-sampling technique (SMOTE) [11] was applied during the model training to improve the performance of the model and avoid overfitting. SMOTE creates new samples for the minority group of the model ($G_2$) with the same statistical properties. Furthermore, the grid search method was performed to find the optimal combination of hyper-parameters of each model [9], based on a stratified 20-fold cross-validation. The classification analysis performance in this study was based on the average evaluation set performance for 20 runs. The following evaluation metrics were used:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2), Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1\ score = \frac{2\ x\ Precision\ x\ Recall}{Precision + Recall} \quad (4)$$

where, TP and TN denote the number of true positive and true negative instances that are correctly classified, and FP and FN indicate the number of misclassified false positive and false negative instances, respectively.

Selecting models with high evaluation accuracy, rules were extracted during the model training using the TE2rules algorithm [12], a novel approach to convert a tree ensemble (TE) for binary classification to a rule list (RL). This algorithm is characterized by high fidelity, as it generates rules from leaf nodes of individual trees and captures the interactions between trees of a TE. Rule selection was also performed considering the models with high training accuracy and a minimum sample of rules.

TABLE II. DATA DISTRIBUTION OF THE CLASSIFICATION MODELS

| Data sets | Subjects | RRMS ($G_1$) | PMS ($G_2$) |
|---|---|---|---|
| Initial | 87 | 66 | 21 |
| Training | 70 | 53 | 17 |
| Over-sample training | 124 | 62 | 62 |
| Evaluation | 8 | 4 | 4 |

RRMS: Relapsing-Remitting MS, PMS: Progressive MS.

### E. Argumentation-based reasoning

Gorgias is a framework for structured argumentation that combines learning and reasoning [13]. It involves constructing arguments using a basic argument scheme, which connects a set of premises (conditions describing a scenario) to the claim of the argument (the desired outcome). Within the Gorgias argumentation theory, two types of arguments are formed: object-level arguments and priority arguments. Object-level arguments are literal claims and can support contradictory claims, leading to arguments attacking one another. Priority arguments, on the other hand, express a local preference between arguments and aim to establish relative strength, tightening the attack relation between them. To visualize the rule tabulation as an internal and application-level explanation, Gorgias Cloud [14] was used. The knowledge is represented by describing the application in terms of object-level arguments. The final output of the proposed model produces an explanation in a language that physicians and patients can understand.

## III. RESULTS

### A. Feature extraction and selection

Statistical analysis was performed on the extracted clinical data and brain MRI lesion features, using the analysis of variance (ANOVA) test to identify if there were significant differences ($p < 0.05$) between the subjects with RRMS ($G_1$) and subjects with PMS ($G_2$). The five most discriminant features that were statistically significant are summarized in Table III.

TABLE III. STATISTICALLY SIGNIFICANT FEATURES BETWEEN THE TWO DIFFERENT GROUPS (RRMS VS PMS)

| **Clinical data** |
|---|
| cerebellarFS, slowtongueFS, facialFS, sensoryFS, dysarthriaFS |
| **Brain MRI lesion texture features** |
| contrastNGTDM, varianceFOS, variancesosSGLDM, sumvarianceSGLDM, modeFOS |

FS: Functional Systems, FOS: First-Order Statistics, SGLDM: Spatial Grey Level Dependence Matrix, NGTDM: Neighborhood Grey Tone Difference Matrix.

### *B. Classification analysis*

The models were trained and evaluated for the selected clinical data brain MRI lesion texture features, and the combination of them. Table IV and Table V tabulate the model evaluation results, respectively. It is shown that the clinical data and the brain MRI lesion texture features can be used to differentiate subjects with RRMS and subjects with PMS. Ensemble tree models (RF, GB, and XGB) gave the best performance. It is obvious that the clinical data gave better evaluation results, achieving an accuracy (ACC) of 95% than the brain MRI lesion texture features, achieving an ACC=79% (Table IV). The evaluation results using the combination of clinical data and texture features reached an ACC=91% (Table V).

The models with high evaluation accuracy were selected to extract rules. More specifically, the RF model of clinical data, and the GB model using both clinical data and texture features. RLs were generated from the selected models using the TE2rules algorithm. Rules were also selected according to the training accuracy and the number of rules. Table VI shows an example after the rule extraction and selection of a GB model. The selected rules consisted of the most important features.

TABLE IV. MODEL EVALUATION RESULTS BETWEEN THE TWO DIFFERENT GROUPS (RRMS VS PMS) AVERAGED AT 20 RUNS USING SELECTED CLINICAL DATA/BRAIN MRI LESION TEXTURE FEATURES

| **Models** | **Accuracy** | **Precision** | **Recall** | **F1 score** |
|---|---|---|---|---|
| DT | 0.94/0.71 | 0.94/0.68 | 0.95/0.85 | 0.94/0.75 |
| RF | 0.95/0.71 | 0.93/0.65 | 1.00/0.95 | 0.96/0.77 |
| GB | 0.93/0.72 | 0.89/0.67 | 1.00/0.93 | 0.94/0.77 |
| XGB | 0.88/0.79 | 0.85/0.76 | 0.93/0.93 | 0.88/0.81 |
| kNN | 0.93/0.62 | 0.94/0.62 | 0.93/0.80 | 0.92/0.69 |
| NB | 0.90/0.60 | 0.98/0.58 | 0.82/0.57 | 0.88/0.57 |
| SVM | 0.88/0.74 | 0.96/0.71 | 0.80/0.80 | 0.84/0.74 |

RRMS: Relapsing-Remitting MS, PMS: Progressive MS, DT: Decision Tree, RF: Random Forest, GB: Gradient Boosting, XGB: Extreme Gradient Boosting, kNN: k-Nearest Neighbors, NB: Naïve Bayes, SVM: Support Vector Machines.

TABLE V. MODEL EVALUATION RESULTS BETWEEN THE TWO DIFFERENT GROUPS (RRMS VS PMS) AVERAGED AT 20 RUNS USING SELECTED CLINICAL DATA AND BRAIN MRI LESION TEXTURE FEATURES

| **Models** | **Accuracy** | **Precision** | **Recall** | **F1 score** |
|---|---|---|---|---|
| DT | 0.90 | 0.89 | 0.93 | 0.90 |
| RF | 0.90 | 0.88 | 0.95 | 0.91 |
| GB | 0.91 | 0.91 | 0.95 | 0.92 |
| XGB | 0.91 | 0.86 | 1.00 | 0.92 |
| kNN | 0.89 | 0.94 | 0.85 | 0.87 |
| NB | 0.88 | 0.90 | 0.90 | 0.88 |
| SVM | 0.91 | 0.90 | 0.95 | 0.92 |

See Table IV for the acronyms.

TABLE VI. AN EXAMPLE OF RULE EXTRACTION USING TE2 RULES

| **Rules** | **Group** |
|---|---|
| **IF** (*cerebellarFS* = Normal **OR** SignsOnly **OR** Mild) **AND** (*facialFS* = Normal **OR** SignsOnly) | $G_1$ |
| **IF** (*cerebellarFS* = Moderate **OR** Severe **OR** Loss) **AND** (*contrastNGTDM* = Medium **OR** High) **AND** (*sensoryFS* = Normal **OR** SignsOnly **OR** Mild) | $G_1$ |
| **IF** *slowtongueFS* = Normal | $G_1$ |
| **IF** (*dysarthriaFS* = Mild **OR** Moderate **OR** Severe **OR** Loss) **AND** (*facialFS* = SignsOnly **OR** Mild **OR** Moderate **OR** Severe **OR** Loss) **AND** (*sensoryFS* = Normal **OR** SignsOnly **OR** Mild) **AND** (*slowtongueFS* = SignsOnly **OR** Mild **OR** Moderate **OR** Severe **OR** Loss) | $G_1$ |
| **IF** (*cerebellarFS* = Normal **OR** SignsOnly **OR** Mild) **AND** (*sensoryFS* = Moderate **OR** Severe **OR** Loss) | $G_1$ |
| **IF** (*cerebellarFS* = Moderate **OR** Severe **OR** Loss) **AND** (*sensoryFS* = Moderate **OR** Severe **OR** Loss) | $G_2$ |

FS: Functional Systems, NGTDM: Neighborhood Grey Tone Difference Matrix, $G_1$, $G_2$: Subjects with RRMS and PMS, respectively

### *C. Argumentation-based reasoning*

Applying Gorgias' argumentation theory, object-level and priority arguments were constructed. The selected rules were modified to object-level arguments. Priority arguments were determined by prioritizing the object-level arguments to resolve the conflicting options. It's worth mentioning that the argumentation theory improved the evaluation ACC of the selected GB model from 91% to 95%. The evaluation results before (GB model) and after applying argumentation theory (GB + ARG models) can be found in Table VII. In addition, an example of a scenario of Gorgias' theory is illustrated in Table VIII, providing the input and output of Gorgias Cloud as well as the explanation in a physical language.

TABLE VII. MODEL EVALUATION OF A SELECTED GB MODEL

| **Models** | **Accuracy** | **Precision** | **Recall** | **F1 score** |
|---|---|---|---|---|
| GB | 0.91 | 0.91 | 0.95 | 0.92 |
| GB + ARG | 0.95 | 0.93 | 0.98 | 0.95 |

GB: Gradient Boosting, ARG: Argumentation theory.

TABLE VIII. AN EXAMPLE OF A SCENARIO USING GORGIAS CLOUD

| **input** |
|---|
| *cerebellarFS*(p4, Moderate).<br>*contrastNGTDM*(p4, Medium).<br>*sensoryFS*(p4, Mild). |
| **output** |
| prove([RRMS(p4)])<br>The statement "RRMS(p4)" is supported by:<br>- "*cerebellarFS*(p4, Moderate); *cerebellarFS*(p4, Severe); *cerebellarFS*(p4, Loss)" and "*contrastNGTDM*(p4, Medium); *contrastNGTDM*(p4, High)" and "*sensoryFS*(p4, Normal); *sensoryFS*(p4, SignsOnly); *sensoryFS*(p4, Mild)" |
| The patient p4 is predicted with relapsing-remitting MS as the *cerebellar functions* are Moderate or Severe or Loss and the *contrast* is Medium or High and the *sensory functions* are Normal or with Signs Only or are Mild. |

FS: Functional Systems, RRMS: Relapsing-Remitting MS, NGTDM: Neighborhood Grey Tone Difference Matrix.

## IV. DISCUSSION

The objective of this study was to implement an explainable AI model in the assessment of MS disease based on clinical data and brain MRI lesion texture features. The primary results indicated that the proposed model can classify the subjects with RRMS and subjects with PMS, achieving a high ACC of 95%. According to the feature importance of the models, high-fidelity explanations were provided.

Recent studies investigated the explainability focused on MS disease. More specifically, Basu *et al.* [3] developed multivariate ML models to predict MS disease activity using XGB and applied SHAP methods to identify the predictive covariates for early identification of MS. A large-scale study was used including demographic, neurological, and laboratory measures, as well as MRI assessment. The models achieved a balanced ACC of 80%. The findings showed that the treatment weeks, the new combined unique active lesion count, the new T1 hypointense lesion count, and the age-related MS severity score were the top predictive covariates. Furthermore, Olatunji *et al.* [4] used different ML models and interpreted them utilizing SHAP and LIME methods for early screening of MS. The input data of the models included clinical features, such as demographic and other laboratory measures. The results indicated that Extra Trees (ET) outperformed the rest of the models with an ACC of 95%. The greatest impact on the model's prediction was shown by age, systolic blood pressure (BP), and alkaline phosphatase. In addition, Conti *et al.* [5] proposed an interpretable ML model to predict cortical atrophy in MS using the XGB algorithm and SHAP plots. The predictive model used demographic data and lesion characteristics. The best prediction performances achieved an average $p < 0.002$ and the most important features were the rimless WM lesion volume, patient age, and the intracortical lesion volume. Another study [6], investigated the LRP heatmaps to explain deep-learning models which classify patients with RRMS and patients with PPMS. Moreover, a preliminary work from our group investigated the rule extraction from brain MRI lesion texture features using ML models to assess the disability progression in MS [14]. However, there is no other study found in the literature that has implemented an explainable AI model that provides explanations in the form of rules to assess MS using brain MRI lesion texture features.

## V. CONCLUSION

Neurologists face challenges in providing personalized treatment for MS due to the unclear understanding of its underlying causes. The pivotal aim is to establish an explainable AI model in clinical use to assist the experts in the diagnosis and prognosis of the evolution of the disease.

## REFERENCES


[1] C. E. P. van Munster and B. M. J. Uitdehaag, "Outcome Measures in Clinical Trials for Multiple Sclerosis," *CNS Drugs*, vol. 31, no. 3, pp. 217–236, 2017.

[2] M. Filippi, P. Preziosa, B. L. Banwell, F. Barkhof, *et al.*, "Assessment of lesions on magnetic resonance imaging in multiple sclerosis: practical guidelines," *Brain*, vol. 142, no. 7, pp. 1858–1875, 2019.

[3] S. Basu, A. Munafo, A. F. Ben-Amor, S. Roy, P. Girard, and N. Terranova, "Predicting disease activity in patients with multiple sclerosis: An explainable machine-learning approach in the Mavenclad trials," *CPT Pharmacometrics Syst. Pharmacol.*, vol. 11, no. 7, pp. 843–853, 2022.

[4] S. O. Olatunji, N. Alsheikh, L. Alnajrani, A. Alanazy, *et al.*, "Comprehensible machine-learning-based models for the pre-emptive diagnosis of multiple sclerosis using clinical data: A retrospective study in the eastern province of Saudi Arabia," *Int. J. Environ. Res. Public Health*, vol. 20, no. 5, 2023.

[5] A. Conti, C. A. Treaba, A. Mehndiratta, V. T. Barletta, C. Mainero, and N. Toschi, "An interpretable machine learning model to predict cortical atrophy in multiple sclerosis," *Brain Sci.*, vol. 13, no. 2, p. 198, 2023.

[6] F. Cruciani, L. Brusini, M. Zucchelli, G. R. Pinheiro, *et al.*, "Explainable 3D-CNN for multiple sclerosis patients stratification", in *Pattern Recognit. ICPR Int. Workshops Challenges. Lecture Notes Comput. Sci.*, vol. 12663, Springer Cham., 2021, pp. 103-114.

[7] J. F. Kurtzke, "Rating neurologic impairment in multiple sclerosis: An expanded disability status scale (EDSS)," *Neurology*, vol. 33, no. 11, pp. 1444–1452, 1983.

[8] C. P. Loizou, S. Petroudi, I. Seimenis, M. Pantziaris, and C. S. Pattichis, "Quantitative texture analysis of brain white matter lesions derived from T2-weighted MR images in MS patients with clinically isolated syndrome," *J. Neuroradiol.*, vol. 42, no. 2, pp. 99–114, 2015.

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 108-122, 2013.

[10] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 785–794, 2016.

[11] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," J. *Artif. Intell. Res*., vol.16, pp. 321-357, 2002.

[12] G. R. Lal, X. Chen, and V. Mithal, "TE2Rules: Extracting rule lists from tree ensembles," pp. 1–17, 2022, [Online]. Available: http://arxiv.org/abs/2206.14359.

[13] A. C. Kakas, P. Moraitis, and N. I. Spanoudakis, "GORGIAS: Applying argumentation," *Argument Comput.*, vol. 10, no. 1, pp. 55–81, 2019.

[14] N. I. Spanoudakis, G. Gligoris, A. C. Kakas, and A. Koumi, "Gorgias cloud: On-line explainable argumentation," *Front. Artif. Intell. Appl.*, vol. 353, pp. 371–372, 2022.

[15] A. Nicolaou, C. P. Loizou, M. Pantzaris, A. Kakas, and C. S. Pattichis, "Rule extraction in the assessment of brain MRI lesions in multiple sclerosis: Preliminary findings," in *Comput. Anal. Images Patterns CAIP 2021. Lecture Notes Comput. Sci.*, vol. 13052, N. Tsapatsoulis, A. Panayides, T. Theocharides, A. Lanitis, C. Pattichis, M. Vento, Eds. Springer Cham., 2021, pp. 277-286.